Is there any value in bulk network traces?

Sid Faber
Member of the Technical Staff
CERT/SEI

Is there value in bulk network traces?

Yes.

Any questions?

What problem are you trying to solve?

Trends

- Particular protocols
- Specific applications or use cases

Existence

- When did something come on line?
- Who uses a service?

Resiliency

- How networks react to an event

Education

Let’s try an example.

Hypothesis:

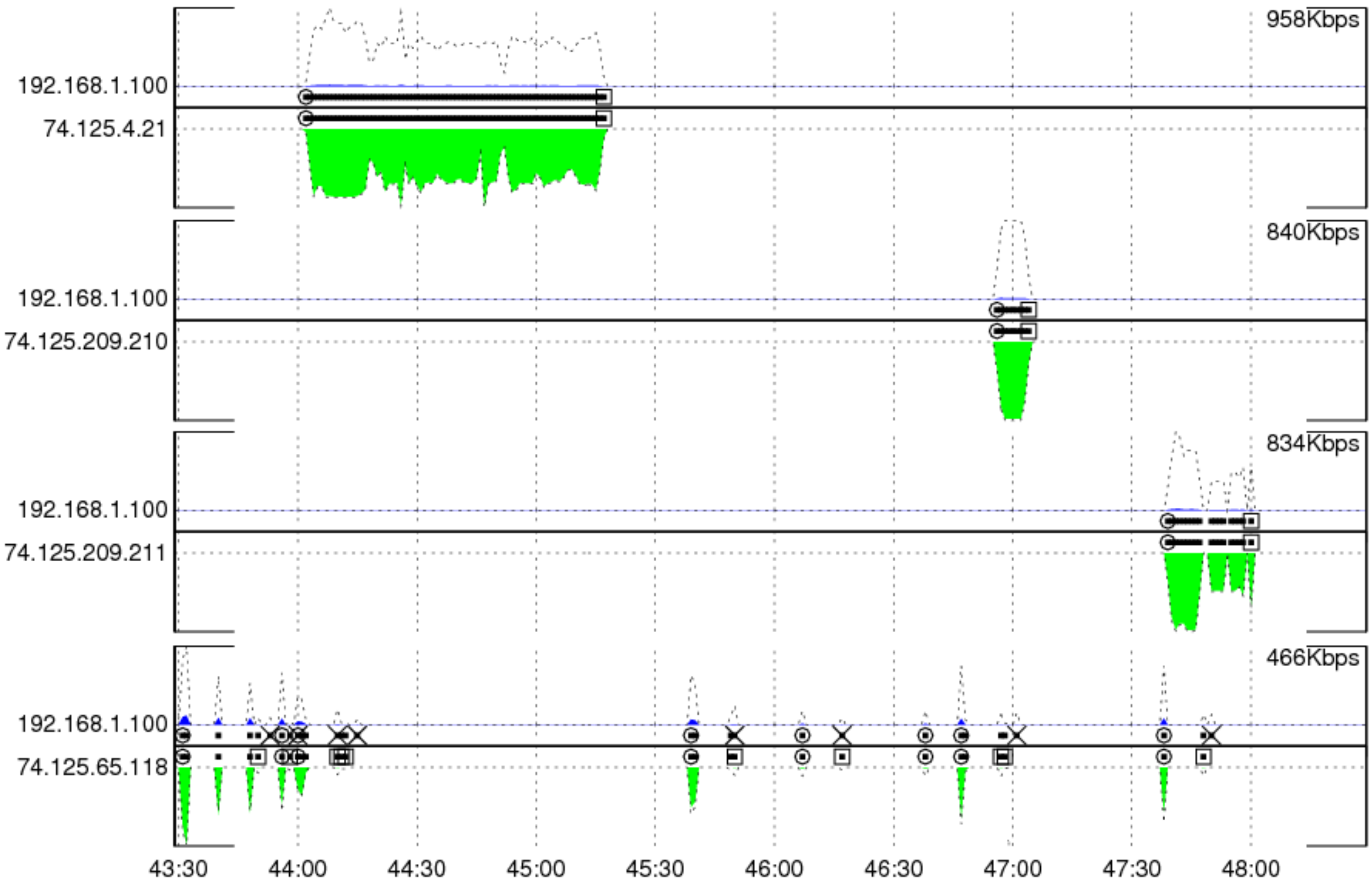
- Internet bandwidth grows by ~40% annually
- Past trends were spurred by audio downloads, then streaming audio, then video clips.
- Now we’re seeing adoption of online TV, and high definition video.

- Is video driving current bandwidth increases? Where are we at on the adoption curve? How will it impact my network?

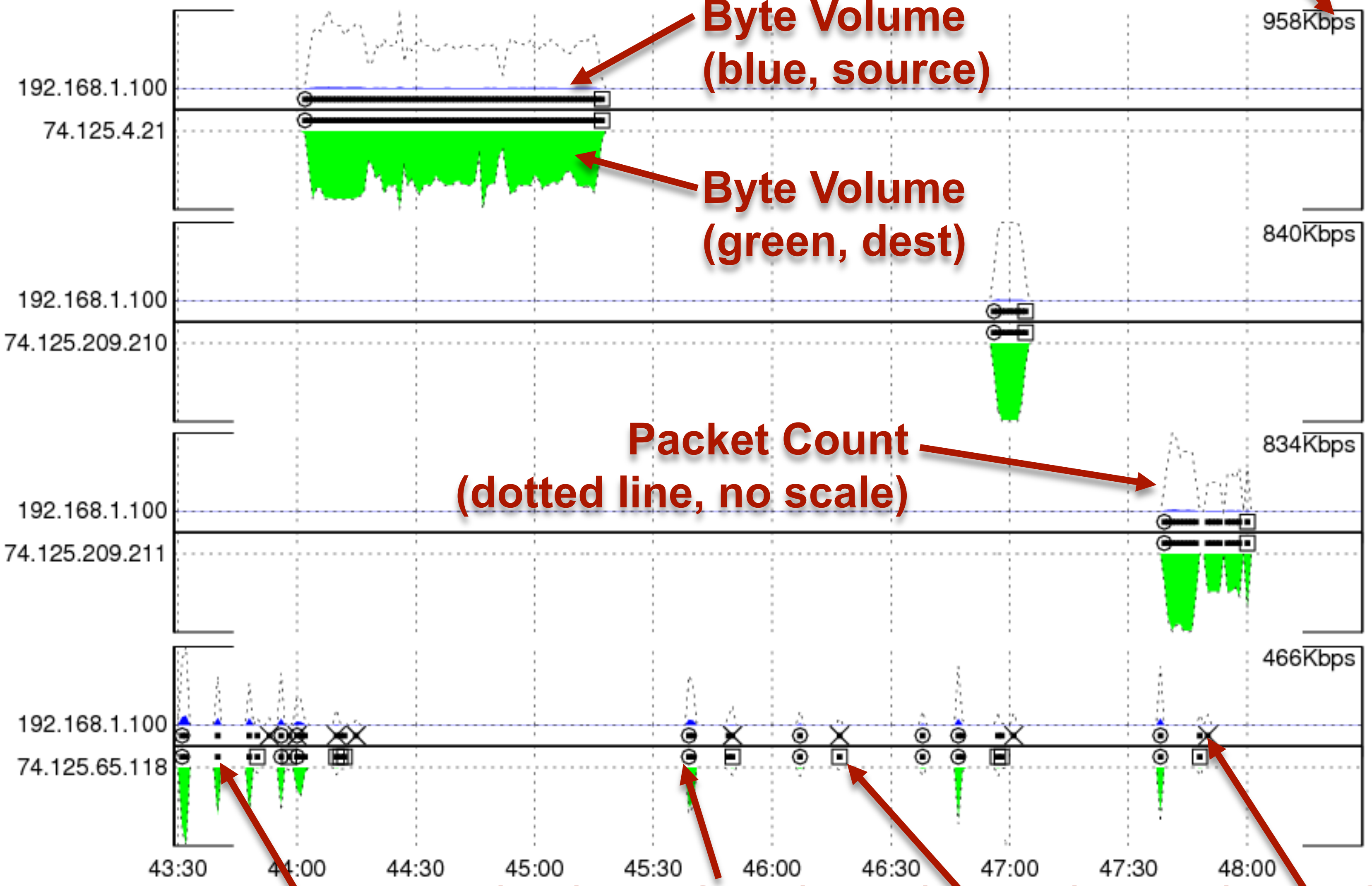
Research plan

- Understand streaming protocols
  - Find features that can identify the protocols
- Look for data to support the research
- Apply the data to the problem

Watch Three YouTube Videos



Watch Three YouTube Videos



Volume Scale

Byte Volume (blue, source)

Byte Volume (green, dest)

Packet Count (dotted line, no scale)

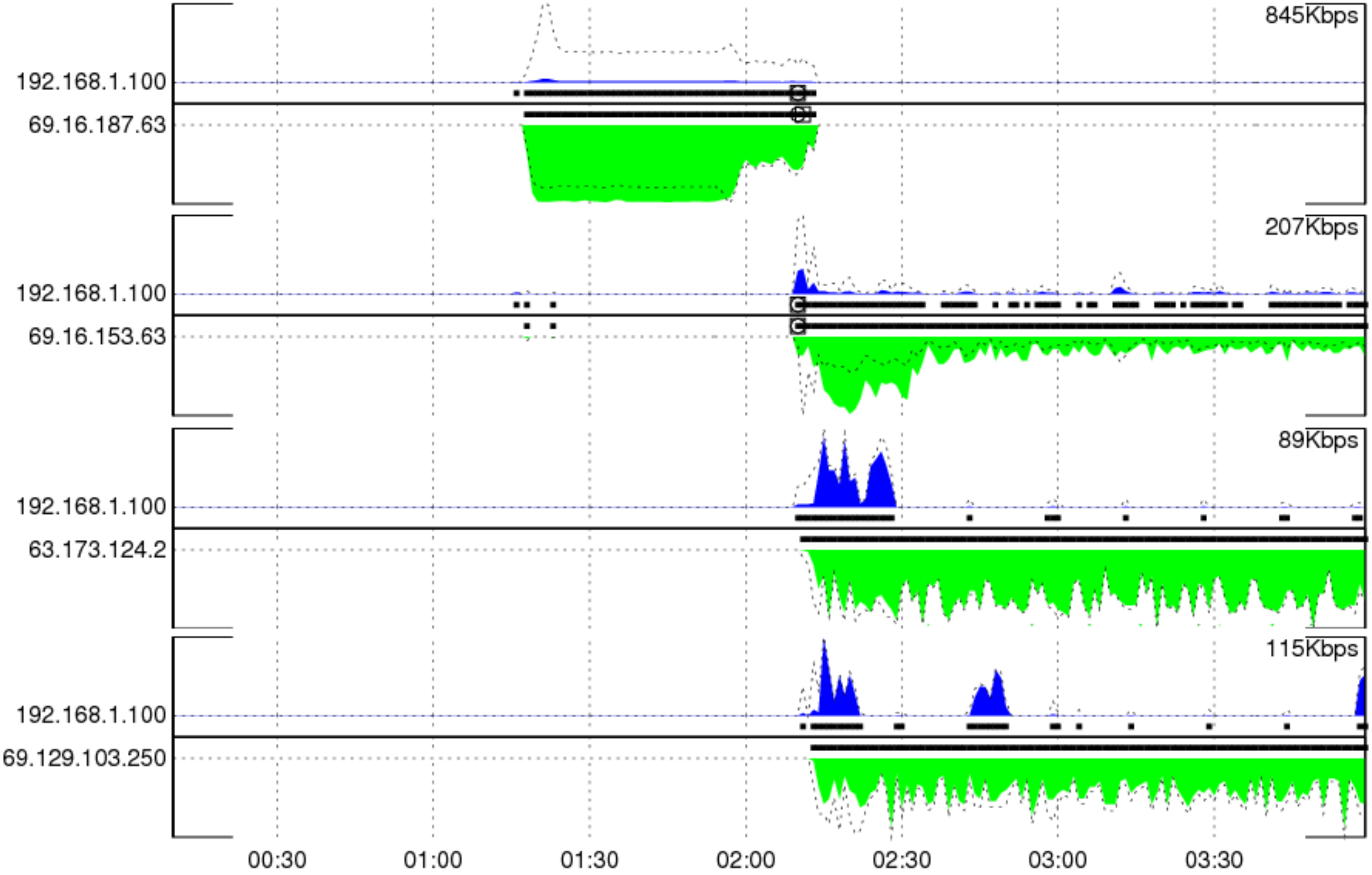
Activity (dot)

SYN (circle)

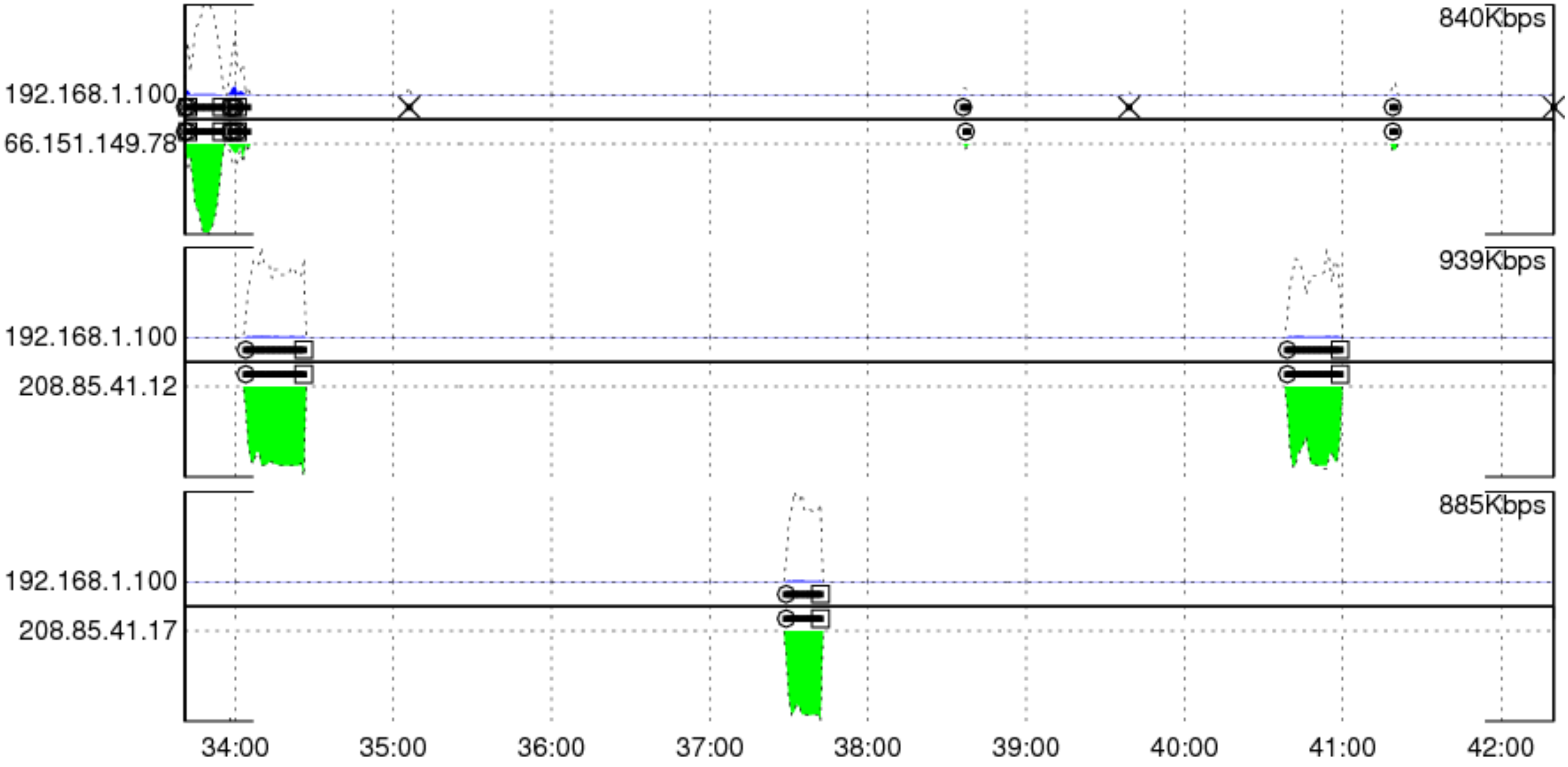
FIN (square)

RST (x)

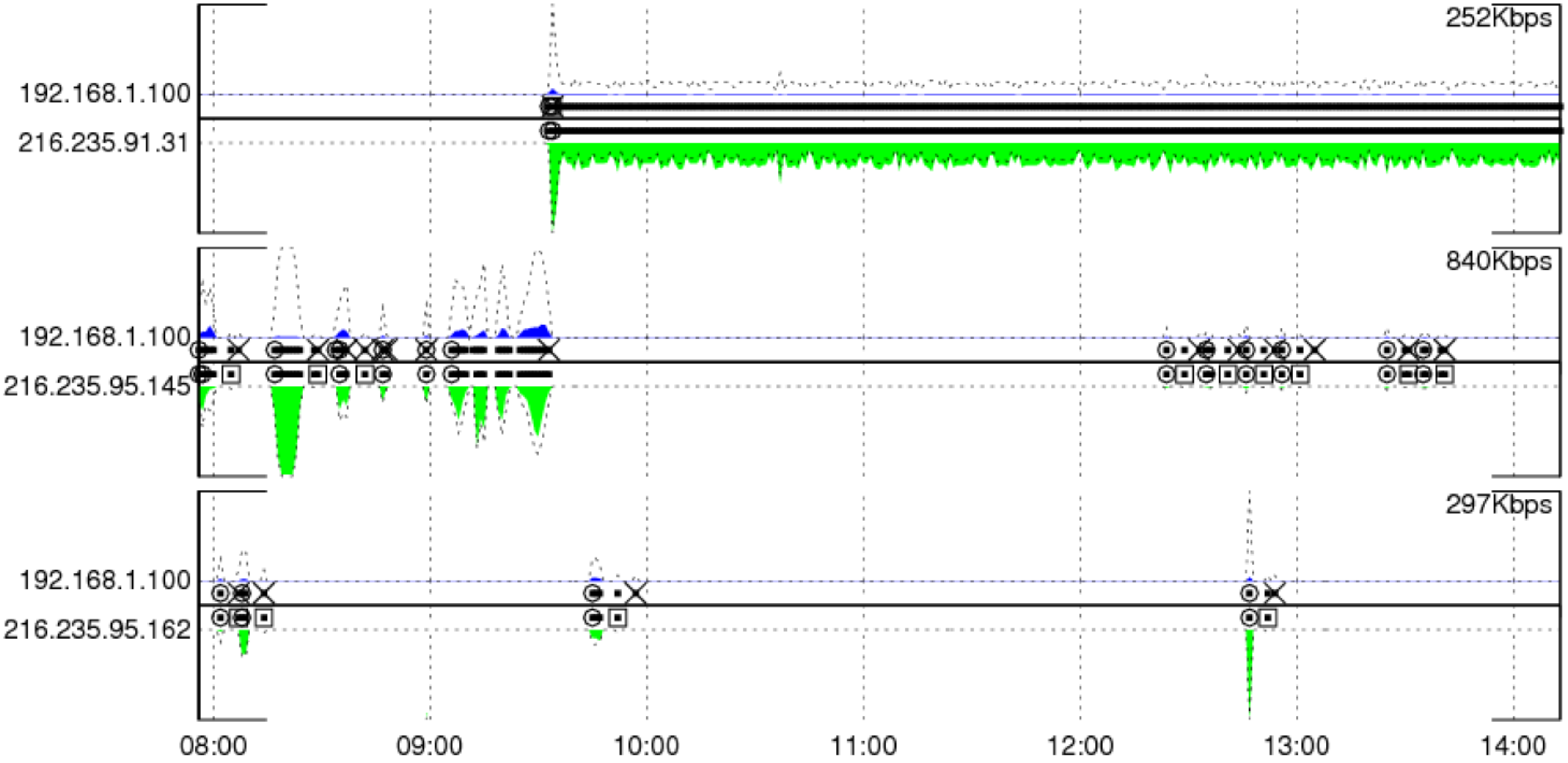
Watch CNN Live



Listen to Three Songs on Pandora



Listen to Live365



Some useful general features

- Overall Bandwidth
- File Delivery protocols vs. Streaming protocols
  - TCP flag patterns
- Use of Content Distribution Networks
- Service port (e.g, HTTP or Shockwave)

Search for data sources

Criteria

- Ongoing data feeds
- Large scale trends across many network types

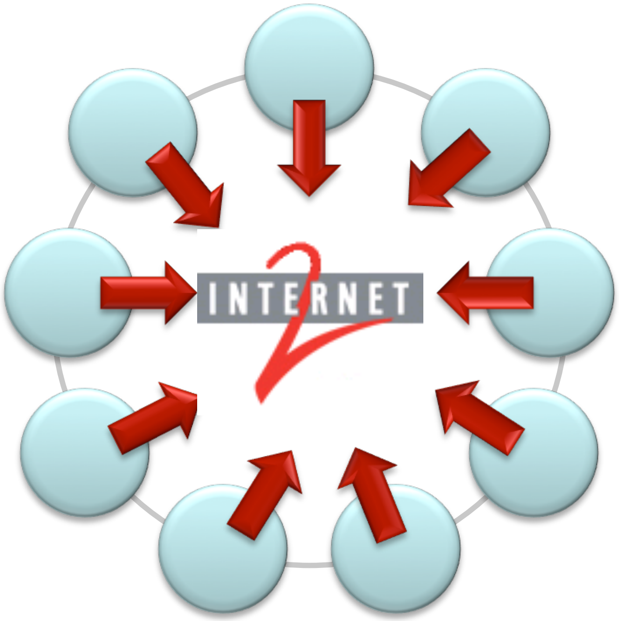
Some Possibilities

- Internet2
- MAWI
- DITL

Data Sources - Internet2

The Internet2 Observatory

- NetFlow v5 in flow-tools format
- Sampled 1:100
- 9 collection points
- Anonymized: lower 11 bits set to 0



http://www.internet2.edu/observatory/archive/proposal-process.html

Data Sources - MAWI

Measurement and Analysis on the WIDE Internet

- Sample point F
- 150Mbps link
- 15 minute snapshot each day
- Unsampled
- Anonymized

Other Data Sources

DITL

Backscatter data

Storm Center Daily Feed

[DatCat]

Challenges: Anonymization

Creates a data silo

Prevents linking in any other IP data sets

- DNS Data
- Geolocation / ownership data
- Blacklists

Not necessarily bad for our research

- Many providers use content distro networks
- Key features are address-independent

Challenges from anonymization are well understood

Challenges: Sampling

It’s often unavoidable

Short term results are unpredictable

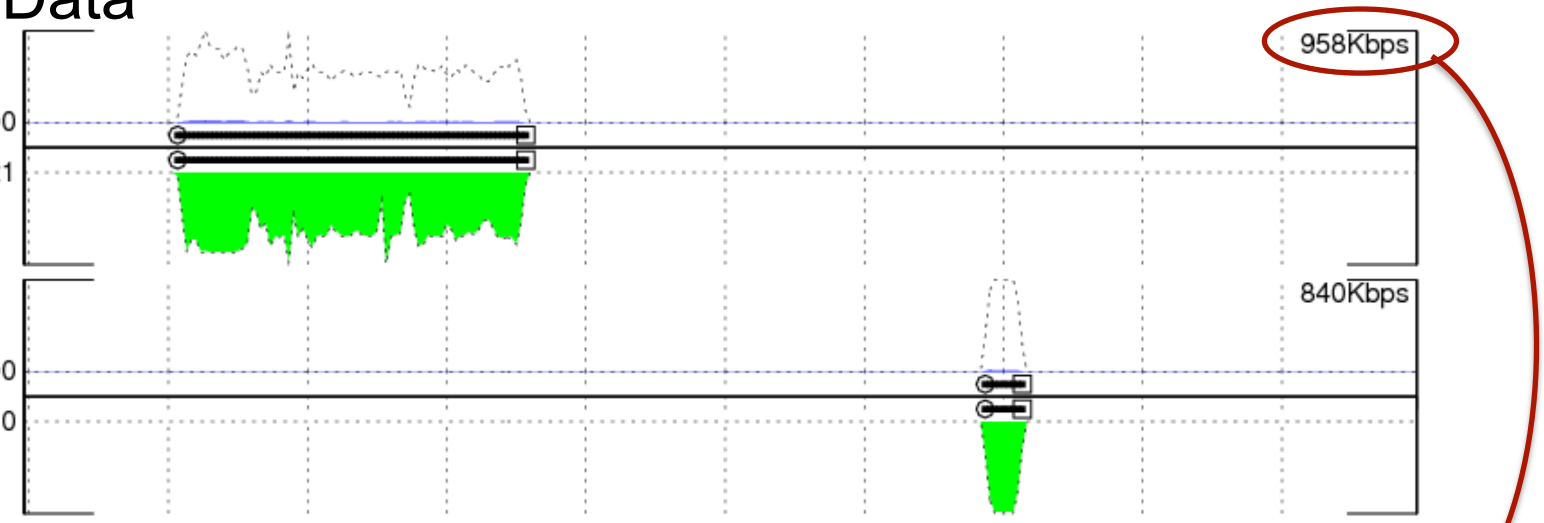
Very significant for our research

- We’re very interested in bandwidth utilization
- Mitigated somewhat because we’re looking at high volumes

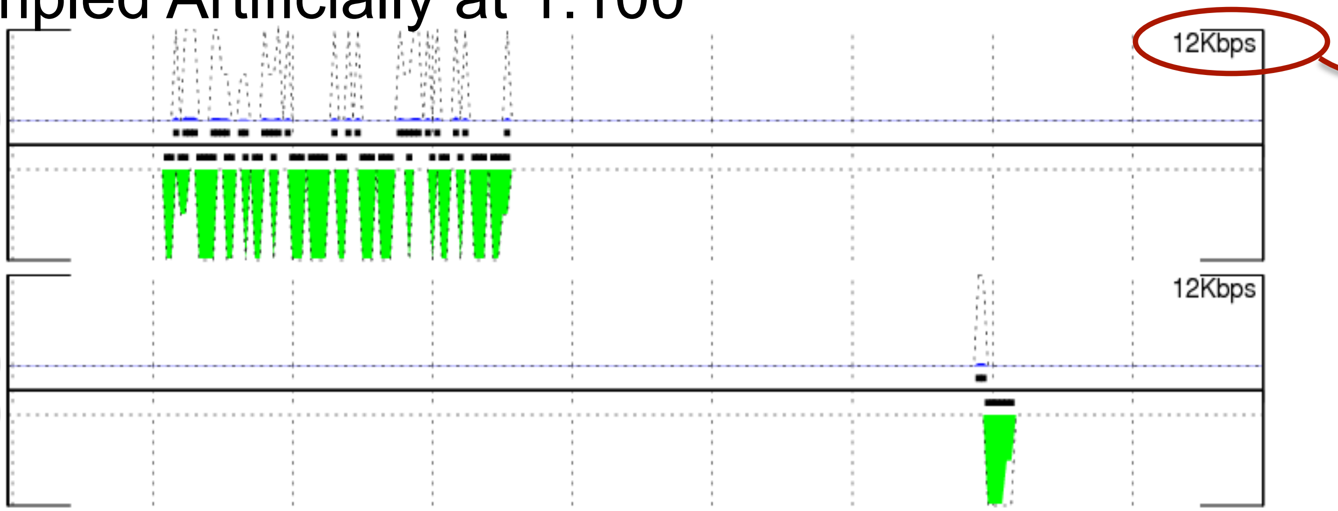
Let’s take a closer look

Watch Three YouTube Videos:

Original Data

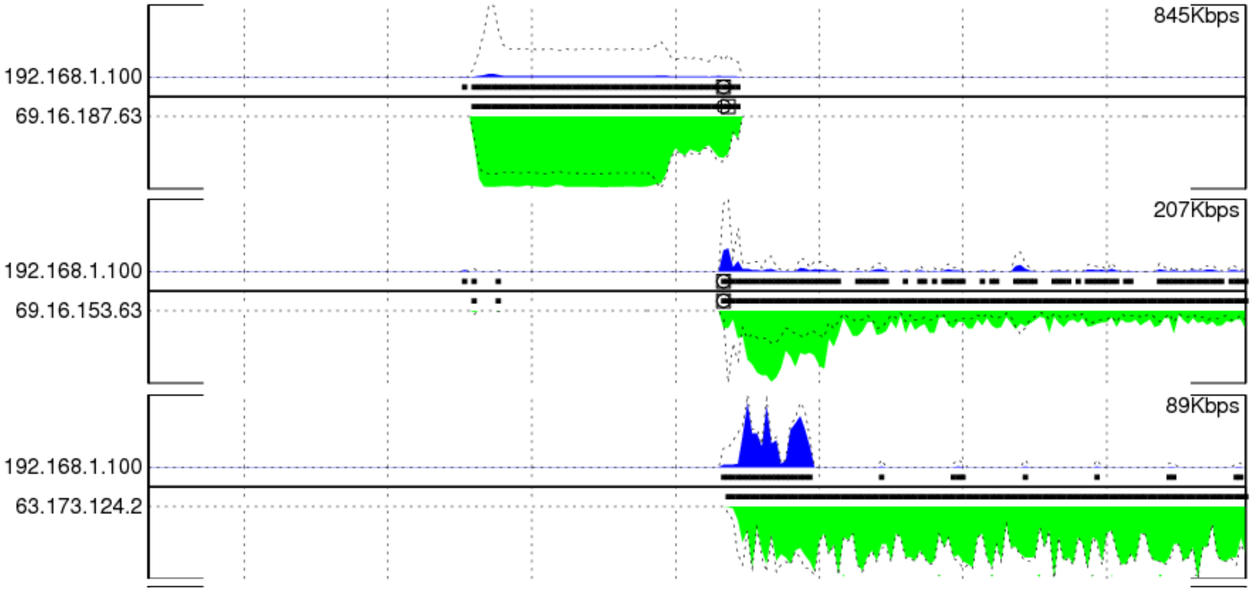


Data Sampled Artificially at 1:100

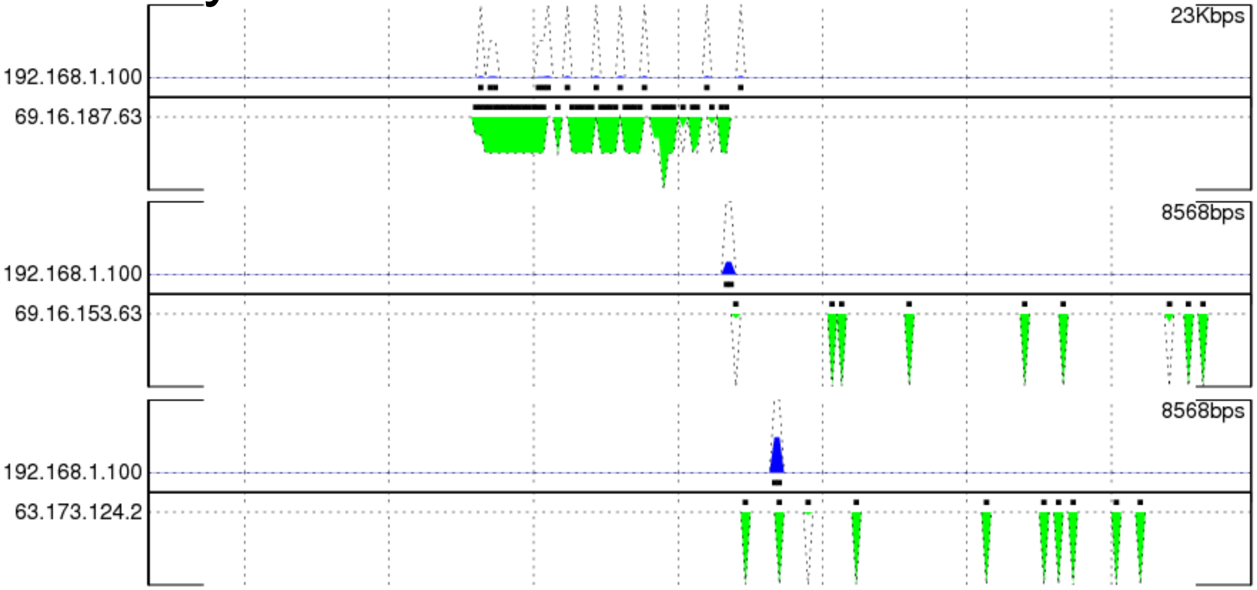


Watch CNN Live

Original Data



Data Sampled Artificially at 1:100

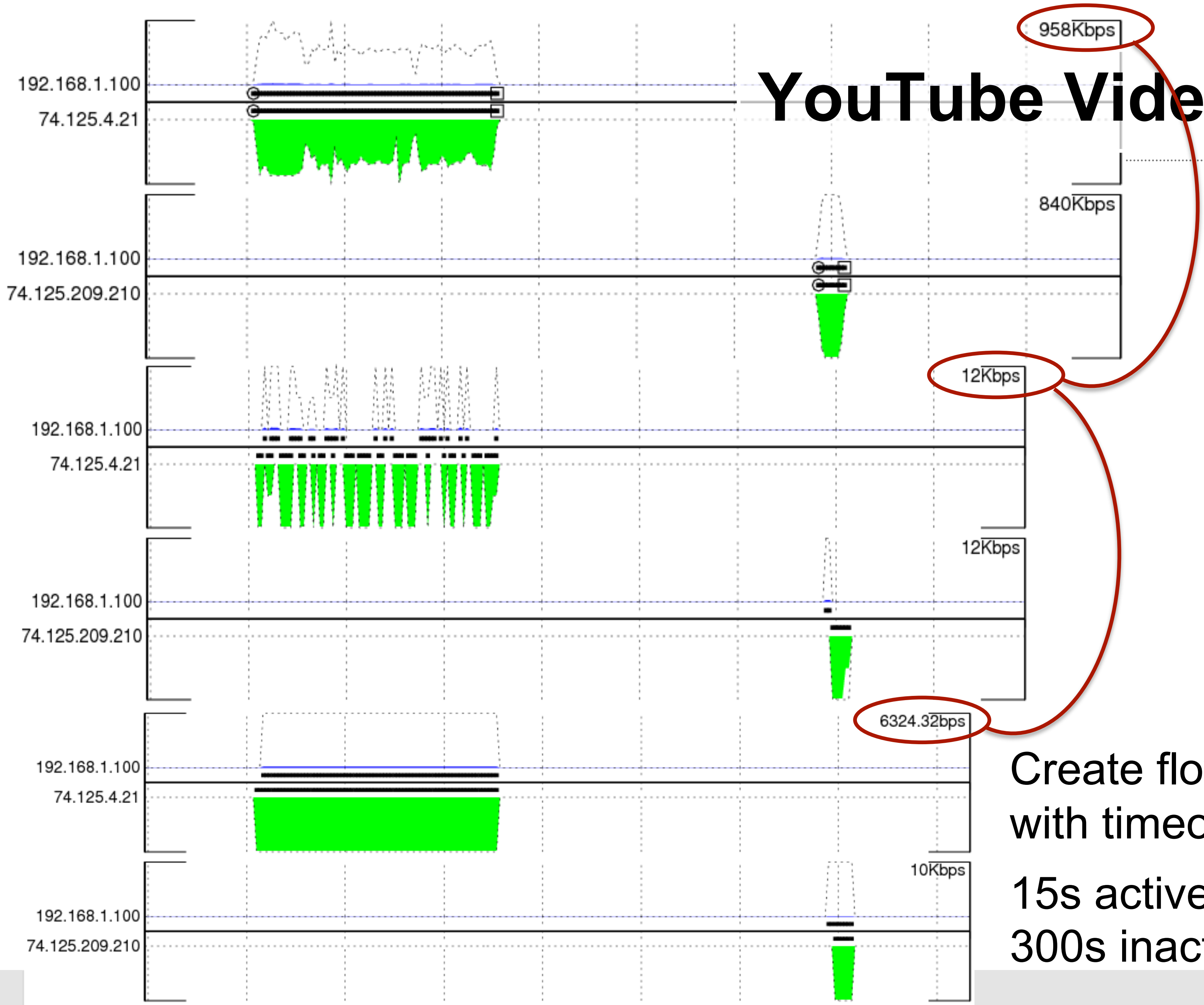


Challenges: Flow

To this point, we’ve been essentially working with packets.

Let’s take a look at the impact of applying flow aggregation and timeouts.

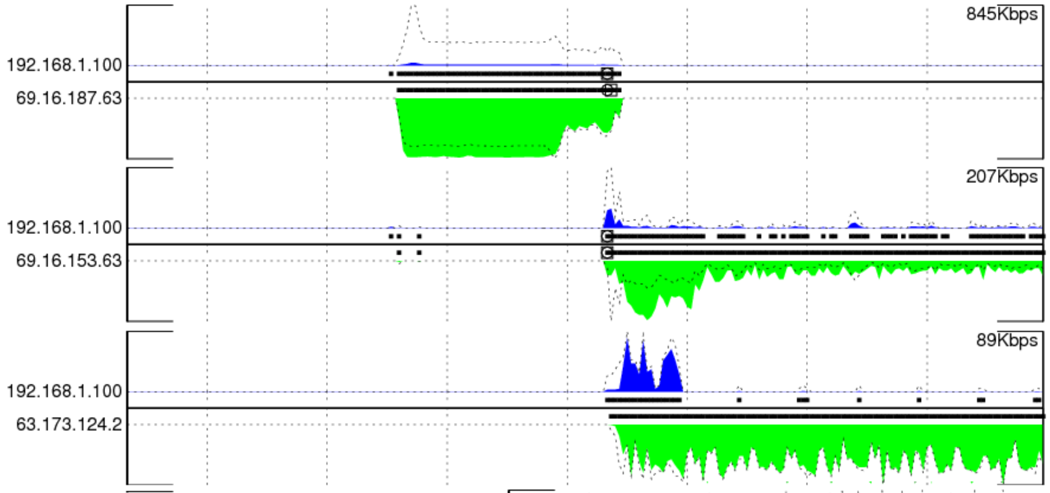
YouTube Videos

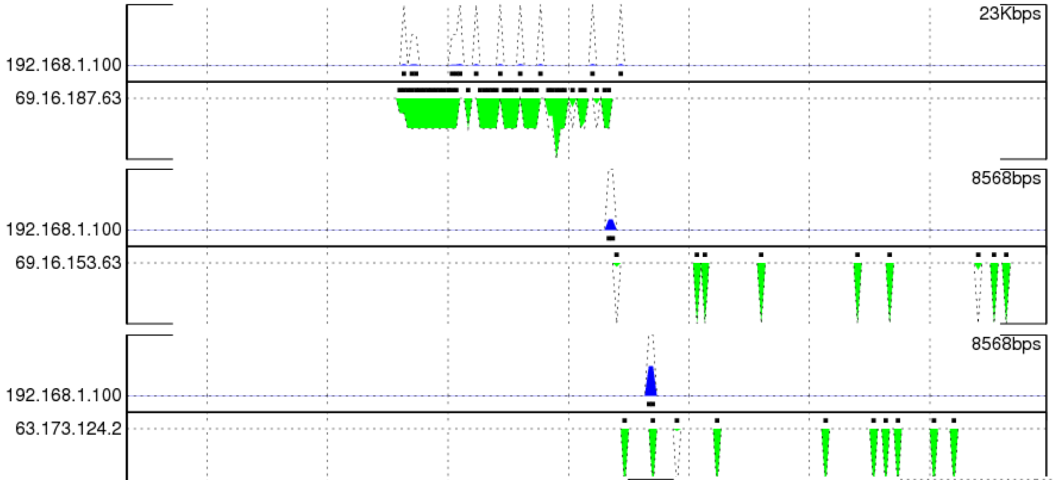


Create flows with timeouts:

15s active
300s inactive

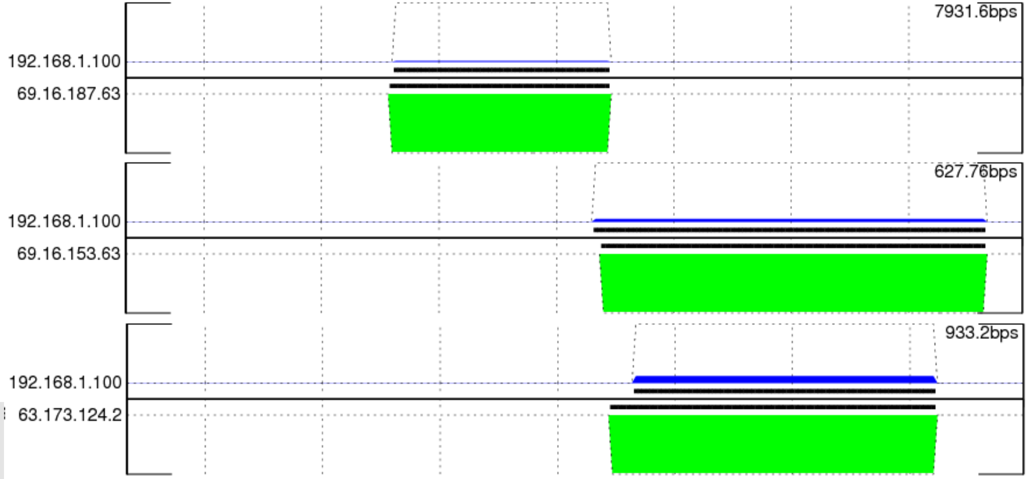
CNN Live





Create flows with timeouts:

15s active
300s inactive



The example, revisited

Is video driving current bandwidth increases? Where are we at on the adoption curve? How will it impact my network?

- We can work around anonymization
- Sampled data makes the problem very challenging
- Working with flow (rather than packets) adds more complexity

Back to the point of the presentation

The question: Is there value in bulk network traces?

The answer: Yes.

A caveat: The data sources have to be tuned to the research

Conclusion

A challenge:

What research do you want to do with bulk network traces?

How can / should we drive bulk network data collection?

Thank You

Sid Faber
sfaber@cert.org